Jurnal Teknologi

Full Paper

EXTREME VALUE ANALYSIS FOR MODELING HIGH PM_{10} LEVEL IN JOHOR BAHRU

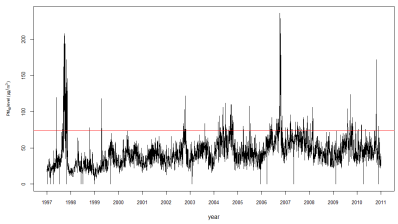
Nor Azrita Mohd Amin[a,b*], Mohd Bakri Adam[a], Ahmad Zaharin Aris[c]

[a]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Perlis, Malaysia
[b]Institute of Mathematical Research, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[c]Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia, Malaysia

Article history
Received
6 June 2014
Received in revised form
18 March 2015
Accepted
1 August 2015

*Corresponding author
norazrita@unimap.edu.my

Graphical abstract



Abstract

Extreme value theory is a very well-known statistical analysis for modeling extreme data in environmental management. The main focus is to compare the generalized extreme value distribution (GEV) and the generalized Pareto distribution (GPD) for modeling extreme data in terms of estimated parameters and return levels. The maximum daily PM_{10} data for Johor Bahru monitoring station based on a 14 years database (1997-2010) were analyzed. It is found that the parameters estimated are more comparable if the extracted numbers of extreme series for both models are much more similar. The 10-years return value for GEV is $104\mu g/m^3$ while for GPD is $289\mu g/m^3$. Based on the threshold choice plot, threshold $u=74$ is chosen and the corresponding 10-years return level is $308\mu g/m^3$. According to the air pollution index in Malaysia, this value is categorized as hazardous.

Keywords: Extreme data, generalized extreme value distribution, generalized pareto distribution, return level, PM_{10}

Abstrak

Teori nilai ekstrem merupakan analisis statistik yang sering digunakan bagi pemodelan data ekstrem untuk pengurusan alam sekitar. Fokus utama adalah untuk membandingkan pemodelan data ekstrem menggunakan taburan nilai ekstrem teritlak (NET) dan taburan Pareto teritlak (TPT) dari segi nilai jangkaan parameter dan aras pulangan. Data maksimum harian bagi PM_{10} untuk stesen Johor Bahru berdasarkan 14 tahun simpanan data (1997-2010) dianalisis. Dapatan menunjukkan nilai anggaran parameter bagi kedua-dua model dapat dibandingkan apabila menggunakan bilangan data extreme yang hampir sama. 10 tahun aras pulangan bagi GEV adalah $104\mu g/m^3$ manakala GPD $289\mu g/m^3$. Berdasarkan plot pilihan ambangan, ambang $u=74$ dipilih dan 10 tahun aras pulangan ialah $308\mu g/m^3$. Nilai ini dikategorikan dalam kategori berbahaya mengikut index pencemaran udara di Malaysia.

Kata kunci: Data ekstrem, Taburan nilai ekstrem teritlak, taburan Pareto teritlak, aras pulangan, PM_{10}

1.0 INTRODUCTION

With the current air quality scenarios, high PM_{10} level is a prominent issue which causes various impacts to human health and material damages. These incidences are a recurring problem in Malaysia especially during haze episodes and in the dry seasons. Improper management of open burning for commercial plantation sectors, heavy industries and business activities has made the situation worse. The region's drier weather conditions have led to escalation in hotspot activities that are caused mainly by land clearing and "slash and burn" agricultural practices. Degradation of ambient air quality standards reduces visibility, impairs air, land and water transportation and seriously affects the Malaysian economy. Prolonged exposure to high concentrations of PM_{10} can be harmful to health especially on eye and throat irritations and respiratory problems among sensitive groups.

Environmental studies and risk assessments deals with atmospheric environments, people and ecosystems. Instead of common events, its main concerns are extreme phenomena (catastrophic). However, most statistical approaches are concerned primarily with the center of statistical distribution or the average value rather than the tail of distribution which contains high observations. Although extreme levels of pollutants concentrations are more worrying, unfortunately not much research investigating these extreme pollutants concentrations based on extreme value concepts are done in Malaysia. Extreme value theory (EVT) has been a standard instrument for many years in the area of forecasting natural catastrophic events as they allow a reliable prediction of the likelihood of uncommon but plausible events to be made. EVT characterizes the tail of distribution and analyzes the extreme data based on generalized extreme value distribution (GEV) and generalized Pareto distribution (GPD) approaches. EVT offers a strong statistical tool for analyzing rare events and predictions of maximum concentrations in certain return periods for air quality management purposes.

EVT provides a concrete theoretical groundwork on which statistical models for describing extreme events are properly set up. General discussions about EVT are about their conventional forms which are Gumbel, Frechet and Weibull that is unified into its general form (GEV) and the recent approach is based on GPD which depends on the threshold value of the data. The major reference on EVT and its applications is referred to Coles [1]. Smith [2] reviewed the statistical techniques for extreme values based on classical methods and threshold approaches. They claim that the GEV methods are undoubtedly easier to apply, because they are based on a single family of distributions for which estimation algorithms are readily available while the GPD methods require more judgments in such matters as how to choose the threshold and how to deal with seasonality. Threshold methods provide more flexibility for the development of alternative models and testing of statistical assumptions but they need large data sets in order to be applied effectively. GPD is capable of extracting observations above a given threshold so that the number of years needed is greatly decreased and the number of samples is increased for extremes whereby this overcomes the major disadvantages of the GEV [3].

Heffernan and Tawn [4] found that the dependence structure exhibits marked seasonality, with extremal dependence between some pollutants being significantly greater than the dependence at non-extreme levels. For extreme value modeling based on threshold techniques, a well-documented issue discussed in Wadsworth and Tawn [5] is the sensitivity of inference from the model to the choice of threshold. They propose a model to account for uncertainty in choice of threshold, under assumptions generated by a penultimate form of EVT and claimed that sensitivity to the threshold is best assessed by examining variation in return level estimates. Bayesian approach is rapidly developing in statistical analysis with the advantage of incorporating expert knowledge or current information to the model inferences. Coles and Powell [6] reviewed literatures that link the Bayesian techniques and extreme value analysis, and used recent advances in Bayesian computational tools for Bayesian extreme value analysis while Coles and Tawn [7] applied the Bayesian extreme framework to the behavior of the rainfall process at extreme levels. Because of its potential to predict the unpredictable, EVT and its methodology are currently in the spotlight. EVT affords some insight into extreme tails and maxima where standard models have proved unreliable [8].

It is important to be aware of extreme events and perform more comprehensive assessments of their consequences. To analyze the extreme observations, EVT plays an important role based on its strong theoretical background particularly in extreme data analysis. In Malaysia, EVT has been used directly and indirectly in some areas such as in hydrology [9], climatology [10] and also in environmental aspects [11, 12]. However these studies are limited to block maxima technique with no further analysis and discussion on threshold based approach. Knowledge and awareness of extreme air pollution are crucial in air quality control. Although there is no prior reason to make assumptions on the probability distribution of air pollutants concentrations, the choice of appropriate statistical distribution models are extremely significant [13]. At present, only a small number of literatures working on distribution fitting in relation to extreme air quality concentrations had been conducted in Malaysia. Hurairah [14] and Yusof [15] worked on extreme value distribution for carbon monoxide and PM_{10} respectively but they were limited to block maxima approach. Sansuddin *et al*. [16] made assumptions on general probability distribution of gamma and log-normal distributions to represent the PM_{10} data for Johor Bahru, Nilai, Kota Kinabalu and Kuantan stations. Unfortunately, only short-term predictions (14 days) for PM_{10} exceedances were

presented and actually nothing much could be planned based on this very short period of time. In addition, the actual concerns in environmental monitoring are future extreme values. It is so that proper preparation and awareness could be made and developed. This means that the usages of mean values are not enough to forecast future extreme cases. This study began with a brief introduction to EVT for GEV and GPD and the associated return levels. Daily maxima PM_{10} data for year 1997 to 2010 were used with the focus on monthly maxima series corresponding to GEV model and threshold exceedances series corresponding to GPD model. Maximum likelihood estimator (MLE) was applied for model inferences and followed by the comparison of these two approaches based on their estimated parameters and return levels.

The main objective of this study is to verify the threshold value in GPD model that gives comparable estimated parameter values and hence to compare the future predictions using both models. Additionally, the return value from GPD model using the threshold chosen based on threshold selection method is determined. This work is expected to contribute some new knowledge into the research area of extreme air quality study in Malaysia in order to better understand, predict, and manage risks of extreme air pollution.

2.0 METHODOLOGY

2.1 Extreme Value Theory

Some references describing major developments in theories and methods for EVT are summarized as follows. Fisher and Tippett [17], the pioneer of the extremal limit theorem introduced three types of limiting distribution of extreme. Fortunately they can be combined into a single family which is known as GEV distribution. Pickands [18] develops the threshold exceedances approaches which are followed by Poisson process and subsequently Pickands [19] introduced the GPD model. The issues discussed in EVT are usually about the correct way of measuring extremes either in terms of number of events that happened or in terms of the size of the extreme events or a combination of them. Most statistical methods are concerned primarily with what goes on in the center of a statistical distribution, and do not pay particular attention to the tails of distribution, or in other words, the most extreme values at either the high or low end. Generally there are two approaches to identifying extremes in real data. The first approach considers the maximum data in periods, for example months or years. These selected observations are also called block maxima. This approach corresponds to GEV distribution. The second approach focuses on the exceedances of a certain high threshold which corresponds to GPD. Details on GEV and GPD can be referred to Coles [1] and Kotz and Nadarajah [20].

Monthly maximum sequence is constructed by choosing the one and only observation available per month. This naturally leads to independent and identically (*iid*) random variables. Threshold method consists of all observations that exceed a suitable threshold. It is obviously unlike the monthly maxima approach since it is not restricted to just one data per month but allows more extreme values to be considered. The only unclear difficulty of GPD model is the issue of selecting the best threshold value. There are some approaches to choosing this value such as Mean Residual Life Plot (MRLP) and Threshold Choice Plot (TCP). However the verification of the accurate threshold values of those methods is not always as simple in practice.

2.2 Generalized Extreme Value Distribution

EVT is based on Extremal Types theorem which state that the limiting distribution of maxima or minima is converging to one of the three distributions called Gumbel, Frechet and Weibull. Then, GEV distribution is a generalization of these three distributions. The cumulative distribution function (cdf) and probability density function (pdf) of the GEV distribution is given by (1) and (2) respectively. The GEV model has three parameters, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$ which refers to location, scale and shape parameter respectively. The ξ value determines the type of GEV distribution. $\xi > 0$ correspond to Frechet distribution, $\xi < 0$ correspond to Weibull distribution and $\xi = 0$ correspond to Gumbel distribution.

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \quad (1)$$

$$g(z) = \frac{1}{\sigma}\exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]-\left(\frac{z-\mu}{\sigma}\right)\right\} \quad (2)$$

Earlier works on EVT take on one of the three distributions and subsequently estimate the corresponding parameters. According to Coles [1], there are two weaknesses regarding this issue. First, a technique is required to choose the most appropriate distribution for the data analyzed and second, the inferences are made with the assumption that the choice is correct. Therefore, a better analysis could be done using GEV where the value of shape parameter, ξ itself will determine the most suitable tail behavior of the data.

2.3 Generalized Pareto Distribution

Data for rare events are often scarce because such events are obviously unusual. Therefore, careful and sophisticated modeling is needed to extract full information from the data and to provide realistic forecasts. Threshold based approach set up threshold values and selects all exceedances. This is totally different from block maxima approach that, although there are other extreme values in a block, only uses the maxima in a series of observations. Therefore GPD

is used as an alternative to make sure all extreme data above the threshold are included for analysis.

GPD distribution could be related to GEV distribution based on the values of their parameters. GEV is developed based on block maxima approach while GPD is based on threshold exceedances approach. Let $y_1,...,y_n$ are *iid* random variables and y is the differences between the observations over the threshold and the threshold itself. The cdf of the GPD distribution defined on $y > 0$ and $\left(1+\frac{\xi y}{\tilde{\sigma}}\right) > 0$ is given by (3),

$$H(y) = 1 - \left(1+\frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \quad (3)$$

$\tilde{\sigma}$ is the scale parameter, ξ is the shape parameter and

$$\tilde{\sigma} = \sigma + \xi(u - \mu) \quad (4)$$

u corresponds to a suitable threshold value while μ and σ are the scale and shape parameter as in GEV model. The pdf is written as in (5)

$$h(y) = \frac{1}{\tilde{\sigma}}\left(1+\frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi - 1} \quad (5)$$

Parameters of GPD are related to parameters in GEV distribution. As in GEV, the value of ξ determine the behavior of the GPD [21]. $\xi < 0$ the distribution of excesses has an upper boundary of $u - \frac{\tilde{\sigma}}{\xi}$, $\xi > 0$ the distribution of excesses has no upper limit and $\xi = 0$ lead to exponential distribution with parameter $1/\tilde{\sigma}$.

Although the ξ values in GEV and GPD should be the same, it is difficult to get this equal value since GEV takes into account the maxima values in blocks while GPD computes the exceedances values from a threshold and this makes the number of samples for both approaches different. However, the ξ value will get closer when the number of samples used for both approaches are almost similar. In determining extreme events using GPD model, the threshold value, u must be chosen appropriately. If it is too low, it will affect the asymptotic fundamental of the model, while if it is too high, it will only generate a few excesses which will lead to high variance in the estimated model. There are a number of approaches available in order to set the threshold value. Basically, we choose as low threshold value as possible depending on a sensible approximation. Coles [1] worked on MRLP which is supposed to be linear above a threshold u_0 to provide a valid approximation to the GPD. MRLP is sometimes difficult to interpret. They also assess the stability of the parameter estimates to a variety of reasonable threshold values. Alternatively, GPD at a range of thresholds is fitted and stability of parameter estimates is sought. If u_0 follows the GPD, then $u > u_0$ also follows the GPD.

2.4 Return Level

Application of EVT in air quality studies are concerned about how well the mathematical theory can be applied to further answer questions relating to the probability that pollutant concentration will exceed a certain level in a given period which refers to the return level. Awareness of the return levels of extreme air pollution events could benefit the development of air pollution risk management practices. Return levels in extremes explain the value of extreme events that occur on average once in a given period. For example, what is the PM_{10} level that will be exceeded on average once in the next 100 years? It is convenient to interpret extreme value models in terms of quantiles or return levels rather than individual parameter values [21]. The return level for GEV with return period 1/p is defined by z_p.

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left\{1 - [-\log(1-p)]^{-\xi}\right\}, & \xi \neq 0 \\ \mu - \sigma\log[-\log(1-p)], & \xi = 0 \end{cases} \quad (6)$$

For GPD model, the return level is explained by x_m in (7) that defines the extreme level that is exceeded on average once every m observations.

$$x_m = \begin{cases} \mu + \frac{\sigma}{\xi}\left[(\zeta_u)^{\xi} - 1\right], & \xi \neq 0 \\ \mu + \sigma\log(m\zeta_u), & \xi = 0 \end{cases} \quad (7)$$

2.5 Model Inferences

Many techniques have been proposed for parameter estimation in extreme value models. These include graphical techniques based on versions of probability plots, moment-based techniques, MLE and Bayesian approach. Each technique has its pros and cons. This study applies the MLE method in estimating parameters of GEV and GPD. The log-likelihood function of the GEV and GPD is given by (8) and (9) respectively.

$$\ell(\theta) = -n\log\sigma - (1/\xi + 1)\sum_{i=1}^{n}\log\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} \quad (8)$$

$$\ell(\theta) = -n\log\tilde{\sigma} - (1/\xi + 1)\sum_{i=1}^{n}\log\left[1+\frac{\xi y}{\tilde{\sigma}}\right] \quad (9)$$

The maximum likelihood estimator, $\hat{\theta}$ of θ is defined as the value of θ that maximize the appropriate likelihood function. For some cases, it is possible to obtain the estimate explicitly, usually by differentiating

the log-likelihood and equating to zero. In more complicated examples, we usually need to use the numerical method approach. The notation of ξ_{GEV} and ξ_{GPD} will be used to represent the shape parameter for GEV and GPD in the subsequent sections.

3.0 DATA DESCRIPTION

3.1 Sampling Site

Johor Bahru is the main city center of Johor in the southern portion of Peninsular Malaysia and is located north of Singapore as shown in Figure 1. It is surrounded by main roads, highly developed industrial and commercial areas, tourist attractions as well as high population density areas. Johor Bahru has a tropical rainforest climate with little variation in temperature and humidity throughout the year. Like other cities in Malaysia, Johor Bahru experiences lots of rain and plenty of heat during most parts of the year. The average annual rainfall is 1778 mm with average temperatures ranging between $25.5\text{-}27.8^{\circ}$C, humidity is between 82 and 86% all year round.

Department of Environment Malaysia (DoE) monitors 52 continuous air quality monitoring stations (CAQM) and 14 different sites of manual air quality monitoring stations to detect any significant change in the air quality level. There are four CAQM in Johor located in Johor Bahru, Pasir Gudang, Muar and Kota Tinggi districts. The air quality monitoring station in Johor Bahru is specifically located at Sekolah Menengah Vokasional Perdagangan Johor Bahru (CA0019) at latitude N $01^{\circ}29.815$ and longitude E $103^{\circ}43.617$ and was established in October 1995. Figure 1 shows the map of air monitoring stations in Johor. This study works on the CA0019 station in Johor Bahru due to the availability of long term data compared to the other stations in Johor.

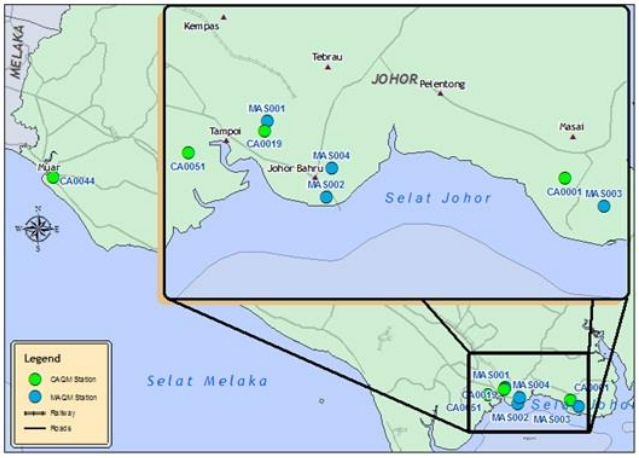


Figure 1 Air monitoring stations in Johor

3.2 The Air Quality Data

The air quality level in Malaysia is described in terms of Air Pollutant Index (API). API is an indicator of the air quality and is developed based on scientific assessment to indicate in a manner that can be easily understood, the presence of pollutants and its impact on health. The API scale and terms used in describing air quality levels are categorized as in Table 1. In extreme value analysis, most of the data considered range between moderate to very unhealthy level for which necessary actions need to be taken. The CAQM measures concentrations of five major pollutants in the ambient air, namely, PM_{10}, sulphur dioxide, nitrogen dioxide, carbon monoxide, and ozone. PM_{10} is used to describe aerosol particles with diameter of less than $10\mu m$ for solids or liquids found suspended in the atmosphere [22]. PM_{10} concentration is related to gases and particulates which are expected to originate mostly from industrial and vehicle emissions and also from some transboundary pollutions involving Malaysia. The three major sources of air pollutions especially in urban areas are mobile (motor vehicles), stationary (power stations, industrial fuel burning process and domestic fuel burning) and the burning of municipal and industrial waste [23].

Table 1 API status indicator

API scale	Air quality status
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very unhealthy
301 and above	Hazardous

The Malaysian guideline on PM_{10} concentration for 24-hours average and 12-months average are $150\mu g/m^3$ and $50\mu g/m^3$ respectively. However, high PM_{10} levels that commonly exceed these guidelines have been a common problem in Malaysia, especially in the dry season. During haze periods, PM_{10} was found as the main pollutant while the other air quality parameters remained within permissible healthy standards [24]. Transboundary haze had contributed to the higher PM_{10} levels recorded intermittently in several areas in Johor in October 2010 due to land and forest fires in the Riau Province in Central Sumatera, Indonesia [25]. Study from Dominick *et al*. [26] mentioned that air pollution in eight selected Malaysian air monitoring stations including Johor Bahru station based on year 2008 to 2009 are predominantly influenced by PM_{10}. In this study, the analyzed data consist of daily maxima PM_{10} data obtained from the DoE from January 1, 1997 to December 31, 2010. The original data were extracted to monthly maxima data to satisfy the independence condition.

4.0 RESULTS AND DISCUSSION

Figure 2 and Figure 3 show the plots of monthly maxima PM_{10} and the daily maxima series with threshold, $u = 74$. There are several obvious high levels of PM_{10} in both series around the end of year 1997, 2006 and 2010. It is due to several episodes of transboundary haze pollution contributed by land and forest fires in Indonesia [27, 25]. In addition, domestic factors in Malaysia especially in urban areas such as industrial activities, vehicle emissions and open burning activities have made the haze situation worse. Block maxima is a classical way to extract extreme series while threshold approach engages with a more natural way of determining whether the observations are extreme or not by allowing all observations greater than a chosen high value into the analysis of GPD. This method is considered as a more efficient use of extreme data but the issue that is always discussed is about the optimum threshold selection. The judgment on the most appropriate threshold value to be used as the basis for generating an optimum threshold exceedances series is an apparent challenge in making sure that the assumption of the GPD model is satisfied [28]. In this study, the GPD model is fitted at a range of different thresholds and the stability of parameter estimates are analyzed. The modified scale parameter, σ^* and ξ_{GPD} against u in the range of thresholds 60 to 100 are plotted in Figure 4 with their confidence intervals. Selecting $u = 74$ allows the estimates to remain near constant. Details on this topic can be referred to Coles [1].

Table 2 provides the summary statistics of the daily maxima, monthly maxima and threshold exceedances series of PM_{10} concentrations for Johor Bahru station. Daily maxima are the original data while monthly maxima and threshold exceedances are the new series extracted from the daily maxima data to satisfy the *iid* assumption of EVT concepts. Daily maxima observations contained 5113 data and the highest recorded PM_{10} concentration was found in the second half of 2006 with $236 \mu g/m^3$ and the mean is $44.18 \mu g/m^3$. The high PM_{10} level is probably due to transboundary pollutions, industrial activities as well as the traffic emissions near the monitoring station. In the monthly maxima series, 168 observations are seen with mean $68.35 \mu g/m^3$ while there are 265 exceedances above threshold $u = 74$ with mean $100 \mu g/m^3$. The mean and the median for threshold exceedances are slightly higher than monthly maxima series indicating that none of the extreme cases are ignored in the analysis of GPD. The skewness is greater than 0 and kurtosis is greater than 3, these imply the existence of extreme values and hence support the usefulness of EVT theory.

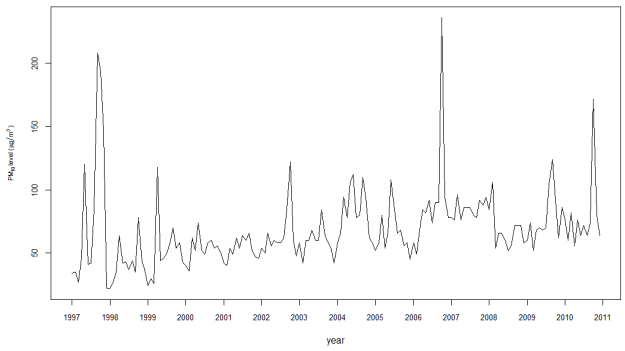


Figure 2 Monthly maxima PM_{10} data in Johor Bahru

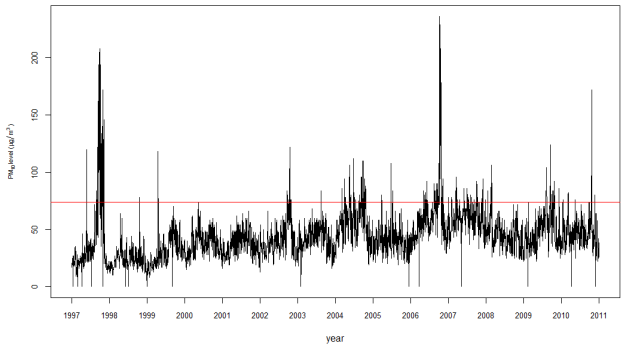


Figure 3 Daily maxima PM_{10} data in Johor Bahru with threshold $u = 74$

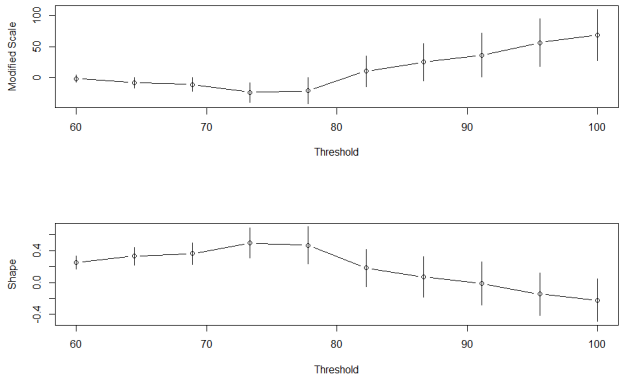


Figure 4 Threshold selection plot for PM_{10} data in Johor Bahru

Maximization of the GEV log-likelihood as in (8) using R software [29-31] for these data leads to the estimation in Table 3. The shape parameter, $\xi_{GEV} > 0$ describes the extracted monthly maxima series as having a Frechet type distribution. Because the attention is on the assessment of estimator values between GEV and GPD, Table 4 summarizes the following details. The first column refers to threshold values within 60 to 100 with a 2 unit increment associated with the number of exceedances in column 2. The third and fourth columns are the estimated values for GPD parameters obtained by maximizing GPD log-likelihood as in (9) with 95% confidence interval. The last column is computed for the scale parameter, $\tilde{\sigma}$ of the GPD based on formula (4) which relates the GEV and GPD scale parameter.

Table 2 Summary statistics of daily maxima, monthly maxima and threshold exceedances of PM_{10} data

	n	Minima	Maxima	Median	Mean	Standard deviation	Skewness	Kurtosis
Daily maxima	5113	8	236	41	44.18	20.33	2.29	15.25
Monthly maxima	168	22	236	62	68.35	2.38	2.38	12.00
Threshold	265	76	236	86	100	1.97	1.97	6.54

Table 3 GEV model fit for monthly maxima PM_{10} data

	Parameter estimates	Standard error	95% confidence interval
μ	55.49	1.70	(52.16,58.83)
σ	20.07	1.23	(17.66,22.47)
ξ	0.06	0.04	(-0.02,0.15)

The bold values in Table 4 refer to the parameter estimates and the confidence intervals of $\tilde{\sigma}$ and ξ_{GPD} if the value of $\tilde{\sigma}$ computed from formula (4) and ξ_{GEV} fall between these two intervals respectively. Theoretically, the ξ_{GEV} and ξ_{GPD} values will be similar at a certain point in the distribution. However, due to the different values of data used in the block maxima versus threshold exceedances approach, the parameter estimation results in different answers. The usage of monthly maxima (12/365=3.8277%) is corresponding to the application of a threshold of 96.7123% on the number of observations and it was found that after this comparable proportion is used, the value for ξ are much more similar. Threshold 78, 80 and 82 give the most similar value for both GEV and GPD parameters, therefore we choose the lowest threshold from these three values which is 78. The quantile plot in Figure 5 and Figure 6 show the validity of the fitted GEV and GPD models for prediction.

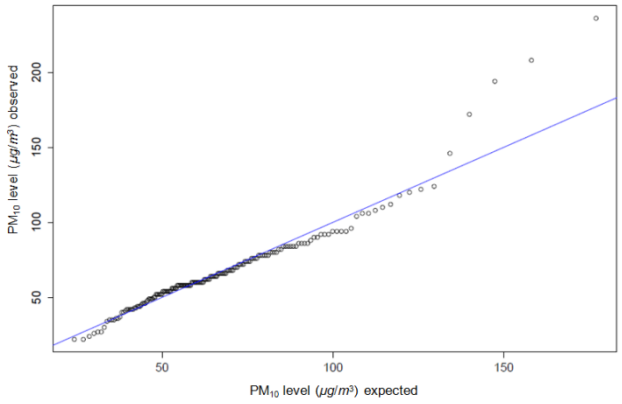


Figure 5 Quantile plot for GEV fit for monthly maxima PM_{10} data

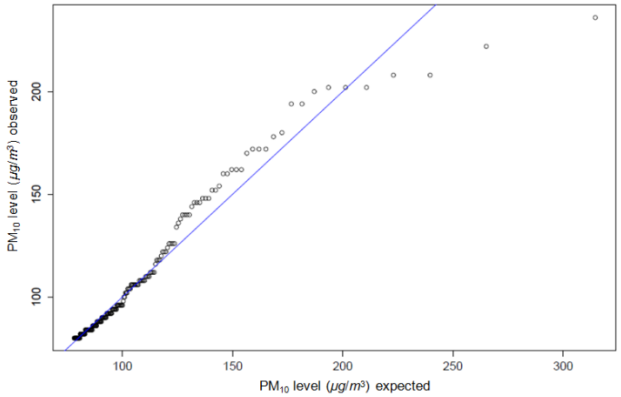


Figure 6 Quantile plot for GPD fit for daily maxima PM_{10} data

The return values for GEV and GPD were computed from the estimated parameter via formula (6) and (7). The 10-years return level for GEV is 104.0269 (95.5282, 112.5255) with s.e 4.33 and GPD is 289.2342 (199.7926, 378.6758) with s.e 45.63. Even though we used comparable estimated values based on the foundation theory, the return values for both models still result in a very large difference in prediction approximation. This is expected since GPD used the exceedances observed above the threshold value which only contain observations greater than $78 \mu g / m^3$ in their model. However in the principal work of GPD, the inferences are based on threshold selection approach. Therefore choosing $u = 74$ produced the estimated parameters $(\tilde{\sigma}, \xi_{GPD}) = (18.82, 0.29)$ and the associated 10-years return level is $308 \mu g / m^3$. This value is reaching the hazardous level and is a warning sign for proper management of future air pollution cases.

5.0 CONCLUSION

Rapid developments in environmental modeling facilities in combination with the theory of statistical extreme values have made advancements in future estimation of possible catastrophic events in a certain return period. Maximum value sampling based on GEV only takes into account one value per block (eg. annual maxima). While in reality the stochastic variability of a maximum is also high in the same year; therefore, one maximum taken for each year is unjustifiable. They may be more than one extreme observation per year and consequently this useful information will be lost. GEV is the classical approach

of extreme environmental modeling while GPD is an alternative method which considers all high values exceeding the threshold. The advantages of GPD in utilizing all extreme observations are very significant in environmental risk study. The results from this study show that the threshold $u = 78$ gives the most comparable estimated parameter for GEV model but the return values show big differences although it was obtained through computation of these comparable estimated parameters. Threshold $u = 74$ is chosen based on the threshold selection method and the corresponding 10-years return value obtained was $308 \mu g / m^3$.

The predicted return levels show that the intensity of future pollution events for PM_{10} will be even worse. Regulation of air pollutants must be properly managed as they bring about harmful effects on human health, vegetation, materials and also on a country's economic developments. In Malaysia, the haze is becoming a predictable annual occurrence, varying only in its severity and duration since 1990 [32]. The worst haze had caused widespread health problems and crisis with losses amounting to billions of ringgits due to disruption of business activities and air transportation. The Malaysian government has taken necessary actions to reduce the occurrence of haze through laws restricting open burning and when necessary cloud seeding is also one of the alternatives used to fight fires to bring down the API to healthy levels. Air pollution from industrial activities could be reduced by enforcing regulations, flue gas treatments, control technologies and careful environmental planning. Open burning should not be carried out indiscriminately especially during the dry and hot season and should be done in accordance with the existing rules and regulations. Although economic development is an important and essential process for every country, at the same time, the prevention of environmental pollution is also of utmost importance.

Table 4 Parameter estimates for GPD model

u	**No. of excesses**	$\tilde{\sigma}$	ξ	$\tilde{\sigma}$ **computed from equation (4)**
60	750	13.80(12.31,15.28)	0.25(0.17,0.34)	20.36
62	642	14.08(12.41,15.75)	0.27(0.18,0.36)	20.48
64	552	14.35(12.48,16.23)	0.29(0.19,0.39)	20.61
66	465	15.55(13.32,17.77)	0.28(0.17,0.40)	20.74
68	401	16.35(13.80,18.90)	0.28(0.16,0.41)	20.86
70	360	15.74(13.04,18.43)	0.33(0.19, 0.47)	20.99
72	311	16.66(13.51,19.81)	0.33(0.17,0.49)	21.12
74	265	**18.82(14.95,22.70)**	0.30(0.12,0.46)	21.24
76	231	**20.71(16.10,25.33)**	0.26(0.08,0.45)	21.37
78	204	**22.59(17.18,28.00)**	**0.23(0.03,0.43)**	21.49
80	181	**24.90(18.53, 31.26)**	**0.19(-0.02, 0.40)**	21.62
82	164	**26.31(19.13,33.49)**	**0.17(-0.06,0.39)**	21.75
84	143	31.81(22.87,40.75)	**0.05(-0.18,0.28)**	21.87
86	132	33.28(23.48,43.08)	**0.03(-0.21,0.27)**	22.00
88	121	35.84(24.91,46.76)	**-0.02(-0.27,0.23)**	22.13
90	111	38.82(26.71, 50.93)	**-0.07(-0.32, 0.18)**	22.25
92	101	43.23(29.73,56.74)	**-0.15(-0.39,0.10)**	22.38
94	93	46.80(32.17,61.44)	-0.21(-0.45,0.04)	22.51
96	84	52.66(36.68,68.65)	-0.29(-0.51,-0.06)	22.63
98	83	49.15(33.35,64.95)	-0.25(-0.50,-0.01)	22.76
100	81	46.66(30.84, 62.49)	-0.22(-0.49, 0.04)	22.89

Note: shape value for GEV parameter is 0.06

Acknowledgement

The author would like to thank the Department of Environment, Malaysia for the data used and the Ministry of Education Malaysia for the SLAI scholarship and research grant (ERGS/1/2013/STG06/UPM/02/3).

References

[1] Coles, S. G. 2001. *An Introduction to Statistical Modeling of Extreme Values.* 3rd ed. London: Springer-Verlag.
[2] Smith, R. L. 1986. *Reliability Engineering*. London: Springer-Verlag.
[3] Yuguo, D., Bingyan, C. and Zhihong, J. 2008. A Newly-discovered GPD-GEV Relationship Together with Comparing Their Models of Extreme Precipitation in Summer. *Advances in Atmospheric Sciences*. 25(3): 507-516.
[4] Heffernan, J. E. and Tawn, J. A. 2004. A Conditional Approach for Multivariate Extreme Values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 66(3): 497-546.
[5] Wadsworth, J. L. and Tawn, J. A. 2012. Likelihood-based Procedures for Threshold Diagnostics and Uncertainty in Extreme Value Modeling. *Journal of the Royal Statistical Society*. 74:543-567.
[6] Coles, S. G. and Powell, E. A. 1996. Bayesian Methods in Extreme Value Modelling: A Review and New Developments. *International Statistical Review*. 64: 119-136.
[7] Coles, S. G. and Tawn, J. A. 1996. A Bayesian Analysis of Extreme Rainfall Data. *Journal of the Royal Statistical Society (Applied Statistics)*. 45: 463-478.
[8] Finkenstadt, B. and Rootzen, H. 2004. *Extreme Values in Finance, Telecommunications and the Environment.* United States of America: Chapman & Hall.
[9] Eli, A., Shaffie, M. and Zin, W. Z. W. 2012. Preliminary Study on Bayesian Extreme Rainfall Analysis: A Case Study of Alor Setar, Kedah, Malaysia. *Sains Malaysiana*. 41: 1403-1410.
[10] Ali, N., Adam, M. B., Ibrahim, N. A. and Daud, I. 2012. Statistical Analysis of Extreme Ozone Data. *Journal of Statistical Modeling and Analytics*. 3: 11-18.
[11] Shaadan, N., Deni, S. M. and Jemain, A. A. 2012. Comparing the severity of PM_{10} using functional descriptive statistics: A case study in Klang Valley. *Journal of Statistical Modeling and Analytics*. 3(1):1-10.
[12] Hasan, H., Radi, N. F. A. and Kassim, S. 2012. Modeling of Extreme Temperature Using Generalized Extreme Value (GEV) Distribution: A Case Study of Penang. *Proceedings of the World Congress on Engineering* Vol 1.
[13] Jiang, X., Deng, S., Liu, N. and Shen, B. 2011. The Statistical Distributions of SO_2, NO_2 and PM_{10} Concentrations in Xi'an, China. *IEEE*. 2206-2212.
[14] Hurairah, A., Ibrahim, N. A., Daud, I. and Haron, K. 2005. An Application of a New Extreme Value Distribution to Air Pollution Data. *Management of Environmental Quality: An International Journal.* 16(1): 17-25.
[15] Yusof, N. F. F. M., Ramli, N. A. and Yahaya, A. S. 2011. Extreme Value Distribution for Prediction of Future PM_{10} Exceedences. *International Journal of Environmental Protection*. 1(4): 28-36.
[16] Sansuddin, N., Ramli, N. A., Yahaya, A. S., Yusof, N. F. F. M., Ghazali, N. A. and Madhoun, W. A. A. 2011. Statistical Analysis of PM_{10} Concentrations at Different Locations in Malaysia. *Environmental Monitoring and Assessment*. 180: 573–588.
[17] Fisher, R. A. and Tippett, L. H. C. 1928. The Frequency Distribution of the Largest and Smallest Member of a Sample. *Proceeding Cambridge Philosophy Society*. 24: 180-190.
[18] Pickands, J. 1971. *Journal of Applied Probability*. 745-756.
[19] Pickands, J. 1975. Statistical Inference using Extreme Order Statistics. *Annals Statistics.* 3: 119-131.
[20] Kotz, S. and Nadarajah, S. 1996. *Extreme Value Distributions Theory and Applications.* London: Imperial College Press.
[21] Gilli, M. and Kellezi, E. 2006. An Application of Extreme Value Theory for Measuring Financial Risk. *Computational Economics*. 1: 1-23.
[22] Juneng, L., Latif, M. T., Tangang, F. T. and Mansor, H. 2009. Spatio-temporal Characteristics of PM_{10} Concentration across Malaysia. *Atmospheric Environment*. 43: 4584-4594.
[23] Sani, S. 1999. *Integrated Environmental Management: Development, Information, and Education In the Asian Pacific Region*. USA: Shiga University.
[24] Payus, C., Abdullah, N. and Sulaiman, N. 2013. Airborne Particulate Matter and Meteorological Interactions during the Haze Period in Malaysia. *International Journal of Environmental Science and Development*. 4(4): 398-402.
[25] Department of Environment Malaysia (DoE). 2010. *Malaysia Environmental Quality Report*. Ministry of Natural Resources and Environment Malaysia.
[26] Dominick, D., Juahir, H., Latif, M. T., Zain, S. M. and Aris, A. Z. 2012. Spatial Assessment of Air Quality Patterns in Malaysia using Multivariate Analysis. *Atmospheric Environment*. 60: 172-181.
[27] Heil, A. and Goldammer, J. G. 2001. Smoke-Haze Pollution: a Review of the 1997 Episode in Southeast Asia. *Reg Environment Change*. 24: 37.
[28] Zin, W. Z. W. 2009. A Comparative Study of Extreme Rainfall in Peninsular Malaysia: with Reference to Partial Duration and Annual Extreme Series. *Sains Malaysiana.* 38 (5): 751-760.
[29] Ribatet, M. A. 2006. R: A User's Guide to the POT Package.
[30] Stephenson, A. G. 2009. R, ismev: An Introduction to Statistical Modeling of Extreme Values.
[31] Stephenson, A. G. 2012. R: Functions for Extreme Value Distributions.
[32] Sinnadurai, J. 2006. Clearing the Air About "The Haze". *Medical Journal Malaysia*. 61(1): 117-121.